Epiwork D2.3: Development of theory and stability analysis for multiscale transportation networks

C. Thiemann, R. Brune, D. Brockmann[1]

[1] MPI-DS, Göttingen, Germany

1 March 2012

Contents

1 Development of theory and stability analysis for multiscale transportation networks

1.1 Goals

One of the challenges of WP2 of the EPIWORK programme is the identification of essential structural features in multiscale mobility and transportation networks as well as social interaction networks that are embedded in the epidemic simulation framework developed in the programme. In large scale networks these essential features are typically masked by the network's topological complexity [17, 20, 23, 2, 18]. Reducing a large-scale network to its core components, filtering redundant information, and extracting essential components are not only critical for efficient network data management relevant to challenges outside of WP2. More importantly, these methods are often required to better understand dynamical processes on networks and the development of order parameters for the predictability of epidemic phenomena which are content of the anticipated deliverable 2.5 of this workpackage.

Although classifications of network elements according to degree, weight, or other centrality measures have been employed in many contexts [5, 15, 29, 26] some of which will be discussed in more detail below, this approach comes with several drawbacks. The qualitative concepts of *hubs* and *highways* suggest a clear-cut, network-intrinsic categorization of elements. However, these centrality measures are typically distributed continuously and generally do not provide a straightforward separation of elements into qualitatively distinct groups. At what precise degree does a node become a hub? At what strength does a link become a highway? Despite significant advances, current state-of-the-art methods rely on system-specific thresholds, comparisons to null models, or imposed topological constraints [24, 21, 23, 28, 19]. In terms of epidemic processes that evolve on these networks, conventional centrality measures do not suggest a clear-cut separation of network elements that are or are not essential for the spreading process. Whether generic heterogeneous networks provide a way to intrinsically segregate elements into qualitatively distinct groups remains an open question. In addition to this fundamental question, centrality thresholding is particularly problematic in heterogeneous networks since key properties of reduced networks can sensitively depend on the chosen threshold. As part of this deliverable we have developed a technique that allows a parameter free identification of essential, spread facilitating network elements. This concept is anticipated to be the foudnation of the development of a systematic quantification of disease spread predictability which is planned for the coming year. In section 1.8 we address these problems and explain the concept of link *salience* that we developed in WP2 [14]. The approach is based on an ensemble of node-specific *perspectives* of the network, and quantifies the extent to which a *consensus* among nodes exists regarding the importance of a link. Link salience is fundamentally different from link betweenness centrality and that it successfully classifies links into distinct groups without external parameters or thresholds. Based on this classification we introduce the high-salience skeleton (HSS) of a network and compute this structure for a variety of mobility and transportation networks.

As part of this deliverable we have performed a comprehensive study of network resilience and developed an approach based on shortest paths in weighted networks and shortest path trees that enables one to investigate resilience of mobility and transportation networks with respect to network element removal. Unlike contemporary methods that fail in densely connected networks (such as conventional percolation theory) this part of the deliverable enables the quantification of modification impact on a single node basis and the identification of sensitive regions of the network. Based on the same foundation we developed the technique of tree dissimilarity which is an essential bridge to dynamic infection hierarchies generated by multiscale dynamics that are in the focus of workpackages 1 and 3.

1.2 Fundamental statistics in multiscale mobility and transportation networks

A key feature that many large-scale human mobility and transportation networks share is their strong structural heterogeneity in terms of link and node statistics and centrality measures. Multiscale transportation networks typically contain a small fraction of hubs characterized by strong connectivity and high centrality scores complemented by a large number of smaller nodes that connect to the hubs. Furthermore, this hub and spokes structure persists accross many spatial scales. Distributions of centrality measures are often scale-free [3, 25].

In order to investigate the extent to which mobility and transportation networks directly related to human traffic share common features or exhibit differences to other transportation networks we devised a comparative analyis of first order statistical networks features in a comparative analysis of the worldwide air transportation network and the global cargo ship network. Our comparative analysis shows that the worldwide air transportation network (WAN) and the global cargo ship network (GCSN) exhibit expected structural properties and, more importantly, their centrality statistics are almost identical, as is illustrated in Fig. 1, that shows the relative frequencies of link weights, node degree, and node flux in both networks.

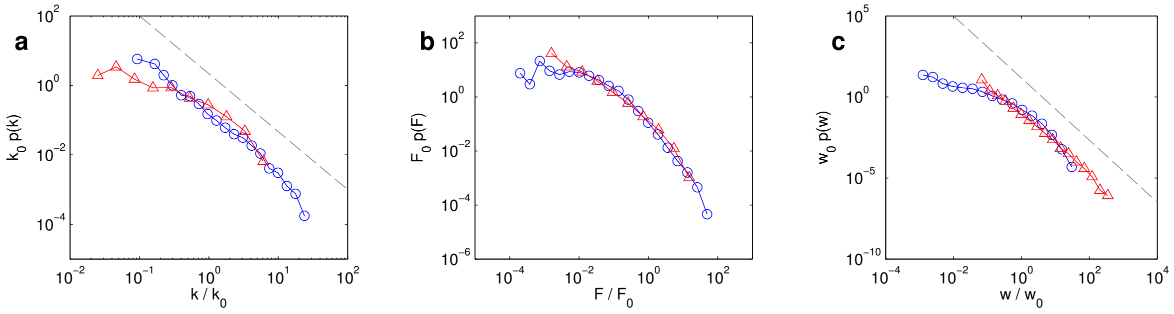

Figure 1: Statistical properties of WAN (blue) and GCSN (red): Panels **a**,**b**,**c** depict the probability density functions of node degree k, node flux F and link weight w. Up to scaling factors these distributions exhibit very similar functional shapes. Link statistics are shown in **d** (link weights w) and **e** (weighted link betweenness b). Approximate scaling behavior is indicated by dashed lines in each panel. Scaling exponents in **a** and **c** are $\alpha = 1.5$ and $\beta = 2$, respectively. The abscissal scaling factors k_0, F_0, w_0 are given by the mean in each distribution.

Figure 1 suggests that w, k and F follow almost identical distributions (up to a scaling factor) and range across many orders of magnitude. Their surprisingly similar shape supports the claim that these networks have evolved according to similar fundamental processes. It has been pointed out [6, 7, 12, 4] that degree, flux, and weight approximately follow power laws based on analyses of smaller, incomplete and regionally focused datasets. This is confirmed for $p(w)$ and $p(k)$, and we find

$$p(k) \sim k^{-\alpha} \quad \text{and} \quad p(w) \sim w^{-\beta} \tag{1}$$

with exponents $\alpha \approx 1.5$ and $\beta \approx 2$. This analysis indicates that human mobility networks on large scales fall into the same statistical category as e.g. te GCSN, which implies that with respect to first order statistical features human mobility networks are not unusual and that general modeling approaches based on first order statistics in situations when data is lacking is to some extent justifiable. This is important for the overall EPIWORK agenda, particularly for filling “network data gaps” in dynamical models for global disease spread.

1.3 Weighted betweenness centrality of links and nodes

Before assessing the stability of networks with respect to modifications, resilience with respect to network failure or anticipated node or link removal, a key goal is to understand and quantify to what extent network elements differ in their role with respect to network function. In this context a commonly investigated measure for link and node centrality is betweenness centrality. The betweenness b of a link (or a node) is the fraction of shortest paths in the entire network of which the link (or node) is part

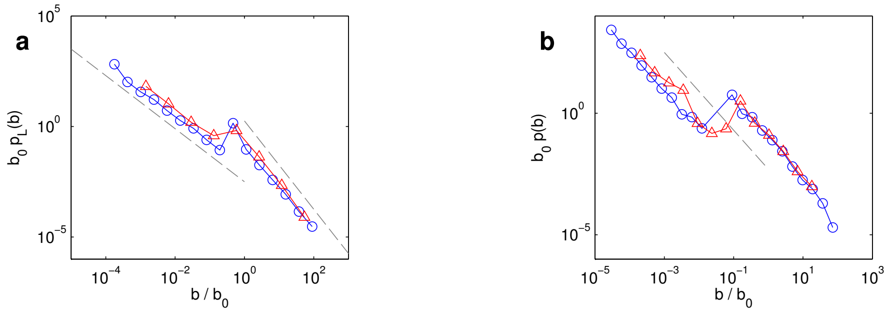

Figure 2: Structure of betweenness centrality of WAN (blue) and GCSN (red): Panels **a** and **b** depict the distributions $p(b)$ of betweenness of links and nodes, respectively. Node betweenness exhibits two distinct regimes separated by a distinct discontinuity for intermediate betweenness values. Each regime is characterized by a scaling exponent $\gamma_1 = 1.6$. Link betweenness exhibits two scaling regimes separated by a marked discontinuity as well, with two different exponents $\gamma_2 = 1.2$ ($b \ll b_0$) and $\gamma_2 = 2.0$ ($b \gg b_0$). The abscissal scaling factors b_0 are given by the mean in each distribution.

of. Betweenness requires the definition of length of a path which in turn requires the definition of length of a link. In weighted networks a plausible choice for the effective length of a link connecting nodes i and j is given by the proximity λ_{ij} defined by

$$\lambda_{ij} = \frac{\langle w \rangle}{w_{ij}}. \tag{2}$$

This notion is particularly useful when considering spreading processes on networks as of prime interest to the EPIWORK agenda. This is so because reciprocal weights provide a natural unit of of inter-event times between nodes of the network. E.g. if w_{ij} represents a per capita tavel rate, it's inverse is the per capita inter-event time. The above definition accounts for the notion that strongly connected nodes are effectively more proximate than nodes that are weakly coupled. The numerator $\langle w \rangle$ sets the typical distance scale $\lambda_0 = 1/\langle w \rangle$ and λ_{ij} is defined relative to it. Based on this effective proximity one can define the length of a path $P(i_0, ..., i_k)$ that starts at node i_0 and terminates at node i_k connecting a sequence of intermediate nodes i_n, $n = 1, ..., k-1$ along direct connections of weights $w_{i_n i_{n+1}}$ by summing of the proximities of each leg in the path. This integrated distance $l(i_0, ...i_k)$ is given by the sum

$$l(i_0, ...i_k) = \sum_{n=0}^{k-1} \lambda_{i_n i_{n+1}} = \sum_{n=0}^{k-1} \frac{\langle w \rangle}{w_{i_n i_{n+1}}}. \tag{3}$$

For a given pair of nodes, i_0 and i_k many paths exists that connect these nodes along intermediate nodes $i_n = 1, ..., k-1$. Using the definition of length of a path above, the shortest path between two nodes is defined as one with minimal l.

$$d(i_0, i_k) = \min_{i_n = 1,..k-1} l(i_0, ..., i_k).$$

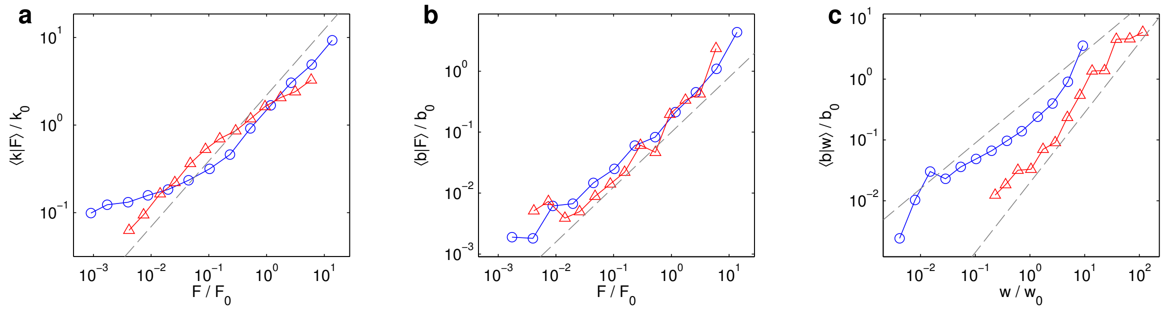

Figure 3: Correlation structure of centrality measures. **a:** The conditional mean degree $\langle k|F\rangle$ as a function of flux F. Both WAN (blue) and GCSN (red) exhibit a similar sub-linear scaling with exponent $\eta \approx 0.75$ across the entire flux range. **b:** Conditional mean node betweenness $\langle b|F\rangle$ as a function of F is almost identical in both networks with almost linear scaling $\zeta \approx 0.9$. **c:** Conditional weight betweenness $\langle b|w\rangle$ as a function of w exhibits different scaling in both networks. The WAN exhibits sub-linear scaling ($\xi \approx 0.75$) contrary to the GCSN for which super-linear scaling is observed ($\xi \approx 1.15$).

We define the effective distance d_{ij} between nodes i and j as this effective length of the shortest path connecting them, i.e. $d_{ij} = d(i,j)$ and denote the unique path associated with it by $P_s(i,j)$ Based on this definition we define the diameter ϕ of the network as the mean shortest-path length over the ensemble of all pairs of nodes. The above notion of shortest paths on weighted networks, is the core concept at the foundation of shortest path tree tomography (Deliverable 2.2).

We computed betweenness centrality b for both, links and nodes based on the set of all shortest paths $P(i,j)$. Figure 2 depicts the distributions $p(b)$ for both networks. Unlike the centrality measures of degree and flux for nodes and weights for links, the distribution of betweenness exhibits a well pronounces discontinuity in both networks. This indicates that in the WAN and GCSN links and nodes segragate into two distinct functional groups. In fact the point b_c at which the discontinuity occurs can be employed to separate links and nodes that belong to the operational backbone of the network[14]. A key orbservation is that both networks exhibit the discontinuity and their distributions of betweenness are very similar, exhibiting scaling behavior in both betweenness regimes.

1.4 Correlations in Centrality Measures

Degree, flux and betweenness typically exhibit positive correlations and scaling relationships with one another. For instance, recently-investigated mobility networks [4, 7, 16] exhibit a sub-linear scaling relation $k \sim F^{\eta}$ with exponent $\eta \approx 0.58$ and $\eta \approx 0.7$. Figure 3 compiles scaling relationships we observe in the WAN and GCSN. To extract the scaling relationship we computed the mean of one centrality measure x

conditioned on a second centrality measure y, that is,

$$\langle x|y\rangle = \frac{\int \mathrm{d}x\, x\, p(x,y)}{\int \mathrm{d}x\, p(x,y)} \tag{4}$$

where $p(x,y)$ is the combined distribution of both. Our analysis shows that both networks exhibit a sub-linear correlation of degree with flux

$$\langle k|F\rangle \sim F^{\eta} \tag{5}$$

with approximately identical exponent $\eta = 0.75$ for both networks and across 4 orders of magnitude of F. This is consistent with previous findings and the intuitive notion that node connectivity increases with traffic. A sub-linear scaling of degree with flux implies that the typical weight $\langle w|F\rangle$ of links connected to nodes of size F scales according to

$$\langle w|F\rangle \sim F/\langle k|F\rangle \sim F^{1-\eta} \tag{6}$$

Since $\eta < 1$ this implies that high flux nodes typically connect to other nodes with stronger links, as expected for transportation networks. The fact that η is almost identical in both networks is additional evidence that similar universal mechanism are responsible for shaping the topological structure of both the WAN and GCSN. Similarly, node betweenness scales as

$$\langle b|F\rangle \sim F^{\zeta} \tag{7}$$

with an exponent $\zeta \approx 1$ in both networks. A linear relationship between node flux and betweenness can be explained by the heuristic argument that typical betweenness values of a node increase linearly with its degree k. Likewise, since shortest paths are computed based on link weights, it is reasonable to assume that node betweenness scales linearly with the typical link weight of a node and thus

$$\langle b|F\rangle \sim \langle k|F\rangle \langle w|F\rangle \sim F^{\eta} F^{1-\eta} = F \tag{8}$$

and hence one expects $\zeta \approx 1$ as observed. Conditional mean of link betweenness as a function of link weight $\langle b|w\rangle$ exhibit approximate scaling. Fig. 3c suggests sub-linear scaling for the WAN as opposed to super-linear scaling for the GCSN.

1.5 Network Resilience, Stability and Control

One of the most fundamental properties of multiscale mobility networks is their resilience towards random failure or anticipated modifications. In particular, in the context of disease mitigation a key goal is the development of network alterations that best slow down proliferation accross the entire networks.

Random failures and targeted modifications are typically investigated using the framework of percolation theory [10, 11]. A random (failure) or selected (anticipated) fraction q of nodes is removed from the network and structural responses of the network

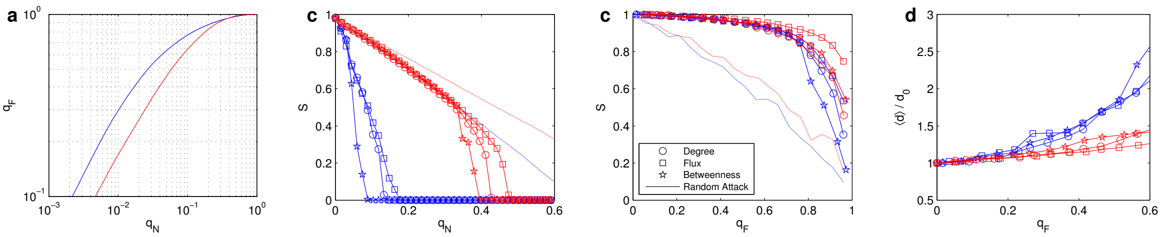

Figure 4: Resilience properties of WAN (blue) and GCSN (red) in response to random and selected node removal: **a)** Fraction q_F of total traffic that is carried by a fraction q_N of top-flux nodes. In both networks a few most central nodes carry substantial amount of traffic in the network. **b)** The relative size of the largest connected component of the networks as a function of the fraction q_N of removed nodes. Nodes are removed according to the rank in terms of degree k (circles), flux F (squares), betweenness b (stars), or randomly (no symbols). Both networks exhibit percolation thresholds under the attack protocols. Numerical values of the percolation thresholds are given in the text. **c)** The largest component as a function of fraction of removed traffic does not exhibit a clear percolation threshold even when almost all the traffic is removed. This indicates that the integrity of the entire network is not altered substantially even if an unrealistically large amount of traffic is reduced. **d)** The response to selected node removal as reflected in network inflation. The panel depicts the diameter of the network as defined by the mean shortest path between all pairs of nodes. Both networks show a substantial increase in diameter as a function of reduced traffic. The effect is more substantial in the WAN.

are investigated as a function of q. Important insight was gained in studies that investigated random or selected node removal in random networks [1, 9, 10, 11]. One of the most important findings of these studies was that scale-free networks with power-law degree distributions respond strikingly different in scenarios that reflect random failures as opposed to selected removal of central nodes. For instance, scale-free networks are relatively immune to random removal of nodes and extremely sensitive to targeted removal of high centrality nodes. Since centrality measures such as degree, betweenness, and flux typically correlate in these networks, this effectively amounts to removal of nodes that function as hubs. One of the essential questions in this context addresses the critical fraction q of removed nodes that are required to disintegrate the global connectivity of the network. This critical value is the percolation threshold q_c: for $q < q_c$ the size of the giant component (the largest subset of nodes that are connected by paths) is typically the size of the entire network. Beyond the percolation threshold ($q > q_c$) the networks falls apart into a family of disconnected, fragmented sub-networks.

The resilience properties of the WAN and GCSN to sequential node removal are depicted in Fig. 4. For each centrality measure (degree, betweenness, and flux), we remove fractions q of nodes according to their rank with respect to k, b, and F, respectively. We compare two different removal protocols. Since both networks are strongly inhomogeneous, removing a fraction of nodes is not equivalent to removing a fraction of traffic (see Fig. 4a). For example 1% of the most connected nodes account for 29.7% of the entire traffic in the WAN and 17.6% in the GCSN, and removal of 10% of nodes with highest flux is equivalent to reducing the total traffic in the WAN by 76.9% and in the GCSN by 64.5%. Because of this pronounced nonlinear relationship, we compare resilience of the network as a function of the fraction of removed nodes q_N as well as the fraction of removed traffic q_F.

Figure 4b depicts the relative size of the giant component S as a function of q_N. Both networks are resilient to random failures (we find that the giant component decreases lineary with the fraction of nodes removed, i.e. $S \approx 1 - q_N$), although the WAN is taking some excess damage from random failures. Furthermore, we observe a percolation threshold for the targeted attacks in both networks. The WAN exhibits a percolation threshold at $q_N^c = 13.8\%$, 17.2% and 9.4%, for node removal according to degree, flux and betweenness. The thresholds are significantly larger for the GCSN at $q_N^c = 44.3\%, 49.0\%$ and 39.5%. In each network the threshold depends only weakly on the choice of centrality measure because of the strong correlation among different centrality measures. Note however that both networks are most susceptible to removal according to betweenness rank, followed by degree and node flux. The overall higher threshold in the GCSN is caused by the greater connectivity σ and mean degree $\langle k \rangle$ of the network (see Table **??**).

Figure 4c depicts S as a function of q_F. The random failures appear to be more effective here because they remove more nodes for a given fraction of removed traffic than the targeted attacks, since not only high-centrality nodes are selected. However, due to the strong nonlinear relationship between q_N and q_F it is evident that both networks are strongly resilient to targeted attacks. Even substantial traffic reduction has virtually no impact on the relative size of the giant component, for instance when 50% of the entire traffic is reduced in both networks, the giant component is still larger than 90% of the original network and no percolation threshold is observed in the range up to 80% of traffic reduction. These traffic reductions are unrealistic when compared to actual perturbations of real transportation networks. Percolation thresholds are therefore never reached under realistic conditions. Another approach that has been applied in unweighted networks [1] is based on the diameter of the network and its response to network disruption. Typically when high centrality nodes are removed from the network, the diameter of the network increases as the shortest paths connecting two arbitrary nodes lengthen due to the increasing lack of hubs that can serve as connecting junctions. Figure 4d shows that this inflation of the network in response to node removal is observed in both networks. This effect is relatively independent of the choice of centrality measure used in the removal protocol. Furthermore, the GCSN is more robust to node reduction, which we believe to be a consequence of the high connectivity of the network. Both, percolation analysis and network inflation have only limited applicability in real world scenarios. Since real world network disruptions never reach the percolation threshold and network inflation only address global structural changes in network properties, a more refined quantity is needed that can determine the response to external perturbations below the percolation threshold and on a node by node basis. In section 1.5 we propose a technique to quantify network resilience that permits the study of network pertubations in a more refined framework and well below the percolation threshold based on shortest path trees. The key idea behind this technique is the ability to quantify the effect of network disruptions for each node and perform network wide statistics of the impact of external pertubations or network disruptions.

1.6 Applied Shortest Path Tree Tomography

As part of deliverable 2.2 we developed the technique of shortest path tree tomography that also represents the computational core of the software package SPATO (www.spato.net) that has been integrated into the GleamViz simulation framework. In order to resolve the limitations of percolation approaches to densely connected human mobility networks we applied the technique to quantify network resilience in regimes well below the percolation threshold.

Although global properties of strongly heterogeneous, multi-scale networks, such as connectivity, clustering coefficient, and diameter, as well as statistical distributions of centrality measures, provide important insight and may serve as quantitative classifiers for networks, they cannot resolve properties and structures on a local scale. On the other hand, local measures such as a node's individual degree, betweenness centrality, or mean link weight of its connections provide local information only and cannot capture global properties. Human mobility and transportation network exhibit important structure on intermediate scales, so it is vital to understand structural properties that are neither local nor global in these networks. One way to approach this is to analyse and investigate the structure of the entire network from the perspective of a chosen node. Clearly, geographic distance is an important parameter in this context as operation costs typically scale with geographic distance. However, in complex multi-scale transportation networks such as the WAN and the GCSN, geographic distance is rarely a good indicator of the effective distance of connected nodes. Typically, high-flux hubs in each network are connected by strong traffic bonds even across very large geographic distances while smaller-flux nodes can be connected by weak links although they may be geographically close.

An alternative representation can be obtained based on the notion of proximity defined by Eq. (2) and effective shortest paths, Eq. (3). Based on this notion we compute the shortest paths of a chosen root node i to all other nodes j. The collection of links contributing to these paths form a shortest-path tree $\mathbf{T}_i$ rooted at i. Spatial representations of such trees are depicted in Fig. 5 for each network and two different root nodes. The radial distance in these figures represents the effective, shortest-path distance d_{ji}. The lines represent the connections of $\mathbf{T}_i$. Note that, although the trees differ in both networks and for different root nodes, high-centrality nodes tend to exhibit the smallest effective (shortest-path) distance to the root node. Note also that the geometry of the networks exhibits significant structural differences in both networks: In the WAN the spatial distribution in the new representation is less regular

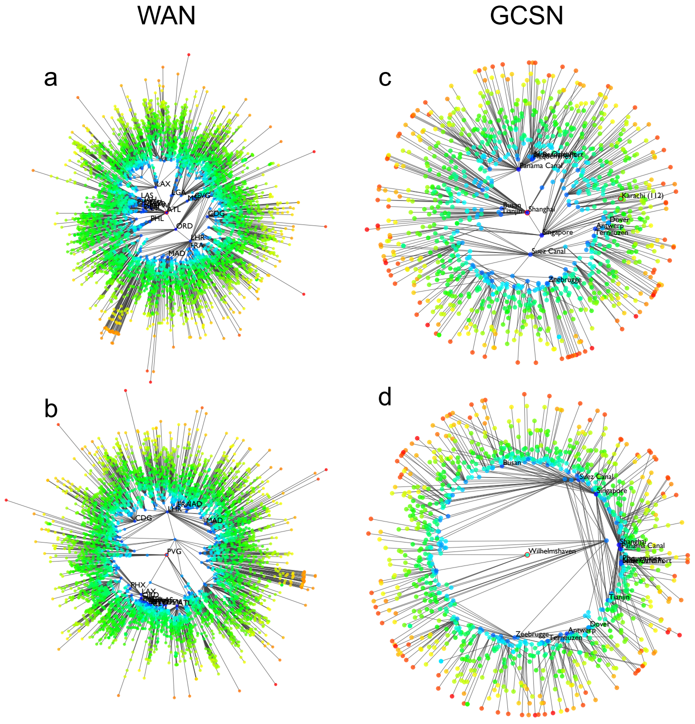

Figure 5: Shortest-path tree structures and effective distances in WAN and GCSN. **Left:** The panels depict the shortest-path trees of airports ATL (Atlanta) and PVG (Shanghai). The radial distance of the remaining airports with respect to these root nodes represent the logarithm of the shortest-path distance to the reference node. **Right:** The panels depict the GCSN shortest-path tree for ports Wilhelmshaven, Germany and Shanghai. Note that the overall structure of both representations is different, yet both networks share the feature of circular arrangement according to node flux, encoded by color (blue represents large flux nodes and orange small node flux). Note that irrespective of the chosen root node, the closest nodes in terms of effective distance are always high flux nodes and small flux nodes are always peripheral in this representation.

and the scatter in effective distance is larger than in the GCSN where nodes reside in a well defined annular region.

In order to understand these qualitative differences and similarities we investigate the distribution of the shortest-path distances conditioned on the type of root node. The results of this analysis are depicted in Fig. 6. Conditioned on the flux of the root node, we compute the distribution of shortest-path distance, that is, $p(d|F)$. Based on this distribution we determine the expected distance of the network from a node with specified flux as

$$\mu_d(F) = \langle d|F\rangle \tag{9}$$

as well as the conditional coefficient of variation:

$$\mathrm{cv}_d(F) = \frac{\sqrt{\langle d^2|F\rangle - \langle d|F\rangle^2}}{\langle d|F\rangle}$$

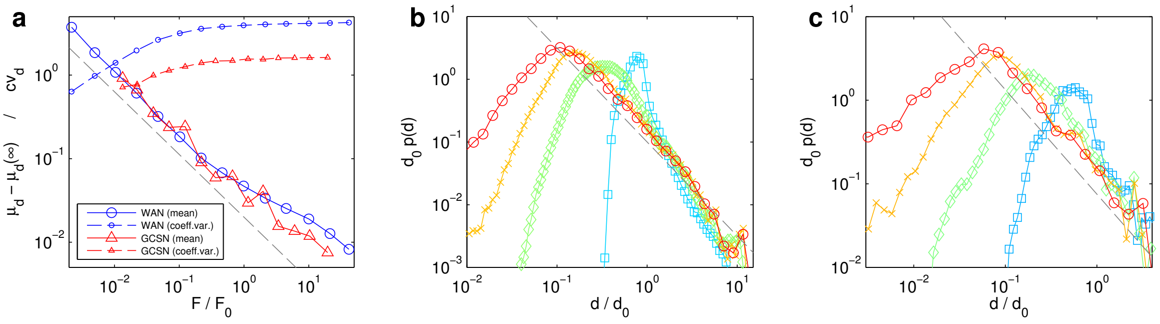

Figure 6: Shortest-path distance statistics. **a:** Conditional mean shortest path (Eq. 9) as a function of node flux as well as the conditional coefficient of variation in shortest paths. Both networks exhibit a universal decrease of expected shortest path with node flux determined by an exponent $\tau \approx 0.75$, see Eq. 10. The variability of shortest-path distance increases in both networks, but less markedly in the GCSN than in the WAN. **b:** The conditional distributions of shortest paths for four subtypes of root nodes (WAN) ranked according to flux, where red markers denote high flux and light blue denote low flux. **c:** Same as in **a** for the GCSN. The dashed grey lines indicate a scaling relation with exponent $\theta \approx 1.5$ (WAN) or $\theta \approx 1.25$ (GCSN).

The quantity $\mu_d(F)$ measures the typical distance from a root node with flux F to the rest of the network. The coefficient of variation measures the statistical variability in d. Figure 6 depicts both quantities for the WAN and GCSN. Note that $\mu_d(F)$ behaves identically for both networks and can described by

$$\mu_d(F) - \mu_d(\infty) \sim (F/F_0)^{-\tau} \tag{10}$$

with $\tau \approx 0.75$. Note that this relation indicates the existance of a lower limit to the typical effective distance for increasing node flux $\mu_d(\infty) > 0$ which implies that even extremely large hubs exhibit a least distance to the rest of the network. Eq. (10) implies that mean effective distance decreases in a systematic way with node centrality and according to the same relation in both networks. However, the coefficient of variation increases monotonically with F, which implies that the variability in effective distance increases with the centrality of the root node. This can also be observed in Fig 6b and c, which depicts the entire distribution $p(d|F)$ for four categories of root nodes of different centrality. For most central nodes $p(d|F)$ increases steeply for small values of d and exhibits an algebraic decay for large distance. As F decreases, $p(d|F)$ attains a sharper peak as small distances disappear from the distribution. This qualitative behavior is observed in both networks. The asymptotic behavior for large effective distances is approximately

$$p(d|F) \sim d^{-\theta}$$

with $\theta \approx 1.5$ for the WAN and $\theta \approx 1.25$ for the GCSN.

A characteristic property of the network representation in Fig. 5 is that regardless of the properties of the root node, the rest of the nodes tend to sort in concentric circles

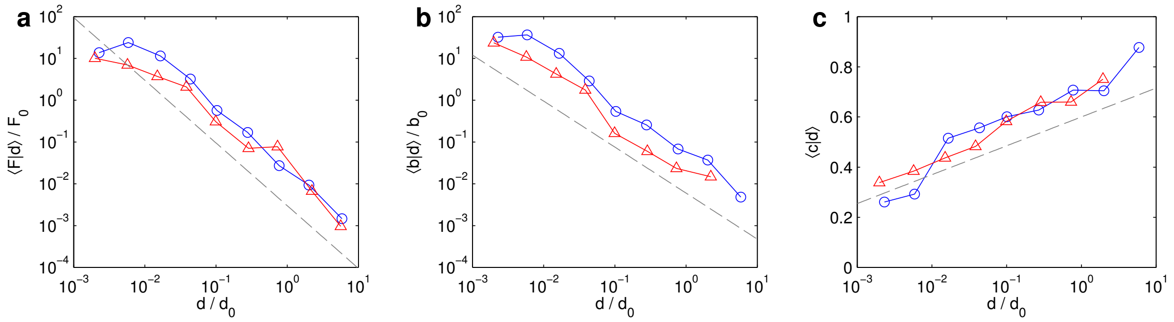

Figure 7: Correlation analysis of shortest paths and centrality measures (**a**) flux, (**b**) weighted node betweenness for WAN (blue) and GCSN (red). Both networks exhibit almost identical scaling relations. The scaling exponents are given by $\omega_F = 1.5$, $\omega_b = 1.1$. **c:** the local clustering coefficient c as a function of effective distance d increases logarithmically.

(effective distances) according to centrality measures. A key question is then how effective distance correlates with centrality measures. If there is a strong correlation between effective distance and node centrality measures, this implies that centrality measures dominate the placement of a node in a network.

In order to determine the relationship between effective distance and centrality measures, we selected 2.5% of the most central nodes, according to degree, flux and betweenness and collected them in a subset of nodes Ω. This fraction of nodes in this set represents 5% of the entire network. The remaining 95% of the nodes are denoted by $\bar{\Omega}$. Based on this subset we determine the distribution $p(x, d|\Omega)$, the probability of finding a node in $\bar{\Omega}$ with centrality measure x (degree, flux, betweenness) and effective distance d to the root nodes in Ω. From this we computed the conditional mean

$$\langle x|d\rangle = \int x\, p(x|d, \Omega) = \int x\, p(x, d|\Omega)/p(d|\Omega), \tag{11}$$

Figure 7 depicts $\langle F|d\rangle$ and $\langle b|d\rangle$ for both networks. Despite their difference, WAN and GCSN exhibit almost identical scaling relations

$$\langle F|d, \Omega\rangle \sim d^{-\omega_F} \quad \text{and} \quad \langle b|d, \Omega\rangle \sim d^{-\omega_b} \tag{12}$$

with $\omega_F \approx 1.5$ and $\omega_b \approx 1.1$, consistent with the intuitive notion that centrality decreases with increasing effective distance from central root nodes. Figure 7 also shows that the local clustering coefficient as a function of d approximately scales according to

$$\langle c|d, \Omega\rangle \sim \log(d/d_0). \tag{13}$$

The logarithmic increase of the clustering coefficient implies that in their peripheral regions the WAN and GCSN become less tree-like. A plausible explanation is that low-centrality nodes that are connected to the root nodes in Ω do not exhibit large fractions of connections among one another, which indicates that high-centrality root nodes function as “feed-in” hubs to low centrality nodes.

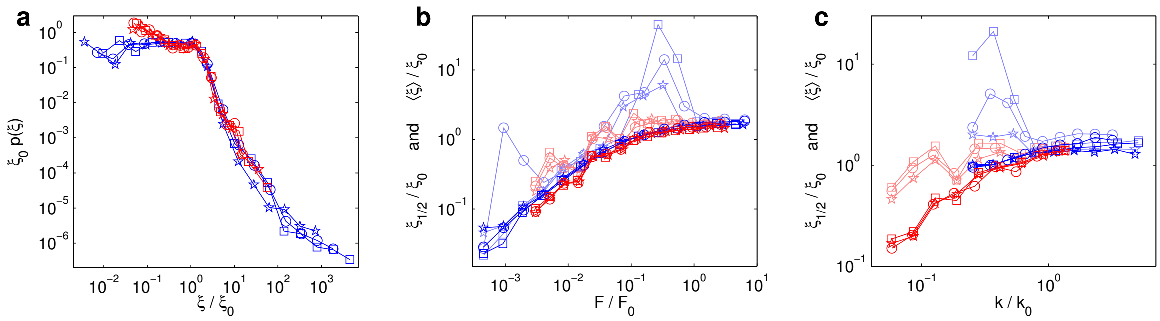

Figure 8: The distribution of impact in response to the removal of central nodes of the system as determined by degree (circles), flux (squares), and betweenness (stars) in the WAN (blue) and GCSN (red). **a:** $p(\xi)$ is the probability distribution of impact factors computed for each node in response to the various attacks. Note that apart from a scaling factor the distribution of impact is identical in all three attack scenarios and both networks. The impact factor ξ ranges over many orders of magnitude. **b:** Universal behavior is also observed in the dependence of impact factor as a function of node flux F. The y-axis measures the normalized median impact ($\xi_{1/2}/\xi_0$, solid lines) and normalized mean impact ($\langle\xi\rangle/\xi_0$, faint lines). **c:** The dependence of impact factor on node degree.

1.7 Resilience and Shortest Paths

The concept of shortest path trees can also give insight into the networks' resilience properties discussed in section 1.5. In response to removal of a fraction q_N of most central nodes or the equivalent fraction of traffic q_F in the entire network, we can compute the impact by investigating the change of shortest-path trees $\mathbf{T}_i$ for each root node i, that is, we can quantify the impact of the network disruption from the perspective of every node. To this end we define a node's impact factor as

$$\xi_i = \frac{\Delta\bar{d}_i}{\bar{d}_i} \tag{14}$$

where $\bar{d}_i$ is the median shortest-path distance from reference node i to all other nodes j, and $\Delta\bar{d}_i$ the change of this median in response to the network disruption. This impact factor is different for every node and the distribution $p(\xi)$ gives insight into the variability of how individual nodes are affected by the network disruption [27]. Figure 8a illustrates $p(\xi)$ for scenarios in which the entire traffic was reduced by 30% through the removal of high-centrality nodes. The distribution $p(\xi)$ is independent of the measure of centrality and also identical in both networks. Below a typical impact of ξ_0 the distribution of impact factors $p(\xi)$ is uniform and for $\xi > \xi_0$ it decreases slowly, ranging over many orders of magnitude. A question that immediately arises is what nodes in the network experience the largest impact. Figures 8b/c depict the mean $\langle\xi\rangle$ and median $\xi_{1/2}$ conditioned on the flux F and degree k. Both the WAN and GCSN exhibit the same dependence, with increasing centrality, the median impact increases monotonically and reaches the typical asymptotic value ξ_0. However, the mean $\langle \xi \rangle$ as a function of F exhibits strong fluctuations for intermediate ranges of F. The explanation for this phenomenon is that the relative impact for low centrality nodes is small because $\bar{d}$ in the unperturbed network is very large. Nodes of intermediate centrality are affected strongly because their mean effective distance to the network $\bar{d}$ is of intermediate size as they primarily connect to hubs in the network by strong links. When the hubs are removed from the network, these nodes experience a strong increase in impact as $\Delta \bar{d}$ is increased substantially. A similar effect is seen in the behavior of $\langle \xi \rangle$ as a function of degree k.

1.8 Link Salience

Human mobility networks such as the worldwide air-transportation network can be represented by a symmetric, weighted $N \times N$ matrix W where N is the number of nodes. Elements $w_{ij} \geq 0$ quantify the coupling strength between nodes i and j. Using the concept of *effective proximity* d_{ij} outlines above effective proximity captures the intuitive notion that strongly (weakly) coupled nodes are close to (distant from) each other [8]. In heterogeneous networks with real-valued weights shortest paths based on this proximity are typically unique.

The central idea of our approach is based on the notion of the *average shortest-path tree* as illustrated in Figure 9a. We define the salience S of a network as

$$S = \langle T \rangle = \frac{1}{N} \sum_{k} T(k) \tag{15}$$

so that S is a linear superposition of all SPTs. According to this definition the element $0 \leq s_{ij} \leq 1$ of the matrix S quantifies the fraction of SPTs the link (i, j) participates in. Since $T(r)$ reflects the set of most efficient paths to the rest of the network from the perspective of the reference node, s_{ij} is a *consensus variable* defined by the ensemble of root nodes. If $s_{ij} = 1$ then link (i, j) is essential for all reference nodes, if $s_{ij} = 0$ the link plays no role and if, say, $s_{ij} = 1/2$ then link (i, j) is important for only half the root nodes. Note that although S is defined as an average across the set of shortest-path trees, it is itself not necessarily a tree and is typically different from known structures such as minimal spanning trees.

1.9 Robust Classification of Links

The most important and surprising feature of link salience is depicted in Figure 9c. For the representative set of transportation networks, we found that the distribution $p(s)$ of link salience exhibits a characteristic bimodal shape on the unit interval. The networks' links naturally accumulate at the range boundaries with a vanishing fraction at intermediate values. Salience thus successfully classifies network links into two

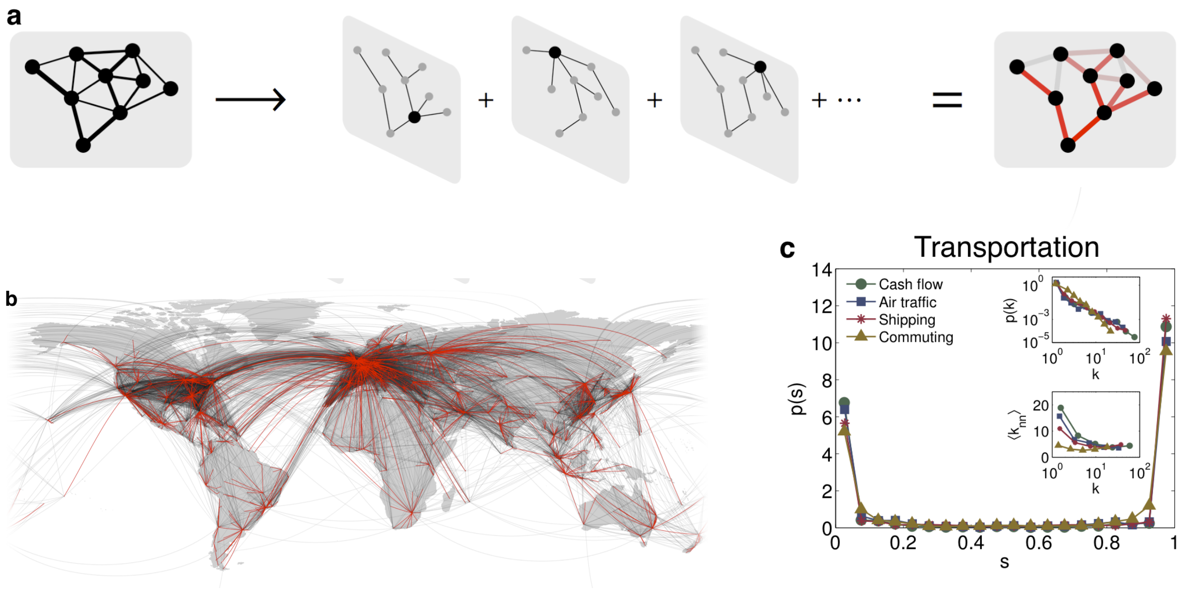

Figure 9: Link salience. (**a**) For each reference node r in the weighted network on the left the shortest-path tree $T(r)$ is computed. The superposition of all trees according to Equation (15) assigns a value s_{ij} to each link in the original network. (**b**) The collection of high salience links (red) for the worldwide air transportation network. The full network is shown in gray. (**c**) The relative frequency of non-zero salience values. The distribution $p(s)$ is bimodal in all networks under consideration (four types of human mobility networks). This key feature of bimodality of $p(s)$ provides a plausible, parameter-insensitive classification of links, salient ($s \approx 1$) vs. non-salient ($s \approx 0$), and implies that nodes in these networks typically *agree* whether a link is essential or not. The high-salience skeleton (HSS) is defined as the collection of links that accumulate near $s \approx 1$. Upper and lower insets depict the degree distribution $p(k)$ of the HSSs and mean next-neighbor degree $\langle k_{nn} | k \rangle$ as a function of degree of a node, respectively. The HSS degree distribution is typically scale-free and the skeletons are typically strongly disassortative.

groups: salient ($s \approx 1$) or non-salient ($s \approx 0$), and the large majority of nodes *agree* on the importance of a given link. Since essentially no links fall into the intermediate regime, the resulting classification is insensitive to an imposed threshold, and is an intrinsic and emergent network property characteristic of a variety of strongly heterogeneous networks. This is fundamentally different from common link centrality measures such as weight or betweenness that possess broad distributions.

The salience as defined by Eq. 15 permits an intuitive definition of a network's skeleton as a structure which incorporates the collection of links that accumulate at $s \approx 1$. Figure 9b depicts the skeleton for the networks.. For all mobility networks considered, only a small fraction of links are part of the high-salience skeleton, and the topological properties of these skeletons are remarkably generic. Note that technically a separation of links into groups according to salience requires the definition of a threshold (e.g. we chose the center of the salience range for convenience). The important feature is that the resulting groups are robust against changes in the value, since almost no links fall into intermediate ranges. Consequently the point of separation is almost arbitrary, yield almost identical skeletons for threhold ranges of 80% of the entire range. One of the common features of these skeletons is their strong disassortativity, irrespective of the assortativity properties of the corresponding original network. Furthermore, all skeletons exhibit a scale-free degree distribution

$$p_{\mathrm{HSS}}(k) \sim k^{-(1+\beta_{\mathrm{HSS}})} \tag{16}$$

with exponents $1.1 \leq \beta_{\mathrm{HSS}} \leq 2.5$. Since only links with $s \approx 1$ are present in the HSS, the degree of a node in the skeleton can be interpreted as the total salience of the node. The collapse onto a common scale-free topology is particularly striking since the original networks range from quasi-planar topologies with small local connectivity (the commuter network) to completely connected networks (worldwide trade). Note that the lowest exponent (weakest tail) is observed for the commuter network, since in a quasi-planar network the maximum number of salient connections is limited by the comparatively small degree of the original network. The scale-free structure of the HSS consequently suggests that networks that possess very different statistical and topological properties and that have evolved in a variety of contexts seem to self-organize into structures that possess a robust, disassortative backbone, despite their typical link redundancy.

Although these properties of link salience are encouraging and suggest novel opportunities for filtering links in complex weighted networks, for understanding hidden core sub-structures, and suggest a new mechanism for defining a network's skeleton, a number of questions need to be addressed and clarified in order for the approach to be viable. First, a possible criticism concerns the definition of salience from shortest-path trees which suggests that s_{ij} can be trivially obtained from link betweenness b_{ij}, for example by means of a non-linear transform. Secondly, a bimodal $p(s)$ may be a trivial consequence of broad weight distributions, if for instance large weights are typically those with $s \approx 1$. Finally, the observed bimodal shape of $p(s)$ could be a property of any non-trivial network topology such as simple random weighted networks. In the following we will address each of these concerns.

1.10 Applications to network dynamical systems

The relevance of link salience to dynamical processes that evolve on networks is a key concern for the EPIWORK programme. Assuming that mobility network connect populations that exchange individuals (metapopulation models) or systems in which nodes represent individuals and links their interaction rate (social network models) contagion phenomena are modeled by transmissions between nodes along the links of the network, where the likelihood of transmission is quantified by the link weights. The central question in this class of models is how the topological properties of the network shape the dynamics of the process (see also deliverable 2.5 and predictability of disease dynamics). Link salience provides useful information about the behavior of epidemic network systems. To illustrate this, we consider a simple stochastic SI epidemic model. At any given point in time, an infected node i can transmit a disease to susceptible nodes at a rate determined by the link weight w_{ji}. We consider an epidemic on a planar disk network. A single node is chosen at random for the outbreak location. At every step of the process each infected node randomly selects a neighbor to infect with probability proportional to the link weight; eventually the entire network is infected. By keeping track of which links were used in the infection process one obtains the infection hierarchy H, a directed tree structure that represents the epidemic pathway through the network. Since the process is stochastic, each realization of the process generates a different infection hierarchy. For different initial outbreak nodes and realizations of the process we calculate an infection frequency h for each link: The number of times that link is used in the infection process, normalized by the number of realizations. The question is, how successfully can link salience, a topological quantity, predict infection frequency h, a dynamic quantity. Figure 10 shows the results for two different link weight scenarios. The top panel shows networks with link weights narrowly and uniformly distributed around a constant value w_0; in the bottom panel link weights are broadly distributed according to a power law. In both cases, link salience is highly correlated with the frequency of a link's appearance in infection hierarchies h, while alternative link centrality measures such as weight and betweenness are not (see Figure 10 insets). The link salience on average gives a much more accurate prediction of the virulence of a link than other available measures of centrality, suggesting that this type of completely deterministic, static analysis could nonetheless play an important role in considering how best to slow spreading processes in real networks.

1.11 Shortest Path Tree Dissimilarity

A more promising approach was introduced recently [22, 13] and is based on the family of shortest-path trees described above. Based on the notion that nodes that are similar in terms of their relation to the rest of the network, a dissimilarity measure is defined based on the difference in shortest path trees of both nodes, that is,

$$D_{ij} = \frac{1}{N} d(\mathbf{T}, \mathbf{T}'). \tag{17}$$

The function $d(\mathbf{T}, \mathbf{T}')$ quantifies the difference or dissimilarity of trees. Choices for this dissimilarity function are for instance parent-dissimilarity or matrix overlap, e.g

$$d(\mathbf{T}, \mathbf{T}') \propto \sum_{nm} (T_{nm} - T'_{nm})^2. \tag{18}$$

For two identical trees dissimilarity vanishes and for trees that share no links it is maximal. If, for example D_{ij} is small among a subgroup of nodes these nodes are likely to be attached to substructures of the network that share a high percentage of features.

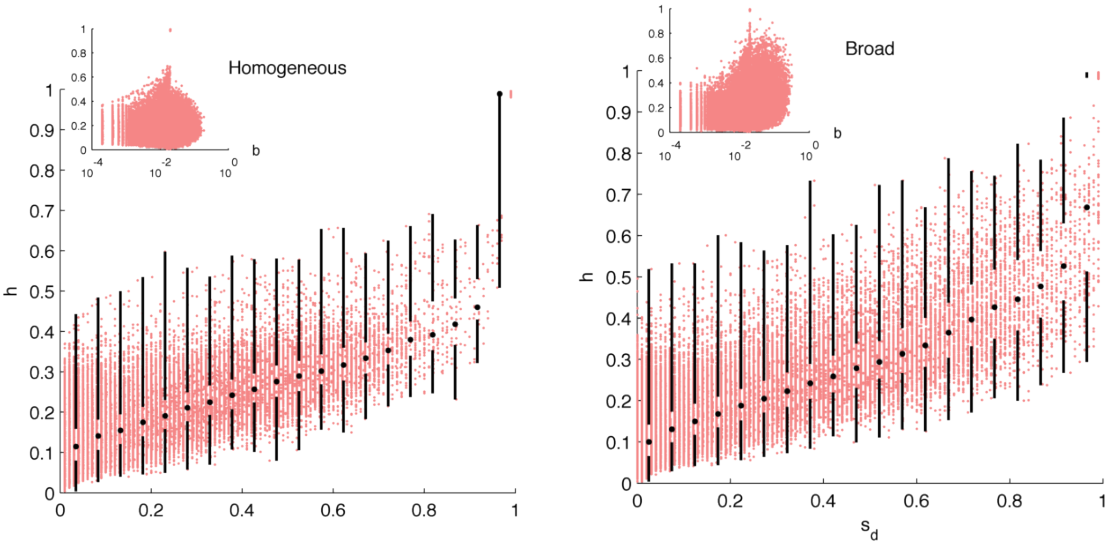

Figure 10: Salience predicts infection pathways in stochastic epidemic models. The scatter plots show the directed salience s_d against the normalized frequency of appearance in infection pathways h for each link in an ensemble of 100 networks, averaged over 1,000 epidemic realizations for each member of the ensemble. The plots are divided horzontally into bins, with the heavy black lines indicating quartiles within each bin. Insets show link betweenness b versus h. **left**, Weights distributed narrowly and uniformly around a constant w_0. **right**, Weights distributed according to $p(w) \sim w^{-(1+\alpha)}$ with $\alpha = 2$.

Tree dissimilarity is not related or correlated with geographic distance nor with effective distance defined by the direct coupling of nodes as defined by Eq. (2). In fact it is typical that nodes can be geographically adjacent or have a strong direct coupling but a large tree dissimilarity. For example airports that serve the same metropolitan area can serve different regions of the network and split the market share of a set of destinations. This induces a large dissimilarity in trees despite close geographic proximity or potentially strong links between the nodes.

Fig. 12 displays the shortest-path trees also shown in Fig 5, except that radial distance is proportional to tree dissimilarities defined in Eq. (17). In this representation, it is clear that the network exhibits substructures that consist of central hubs to which peripheral nodes are connected and that share parts of the entire network. Based on tree dissimilarity we performed a hierarchical clustering of nodes in the WAN and the GCSN splitting the network in 20 groups each such that tree dissimilarity is small within the groups. These communities are labeled with different colors in Fig. 11. We see that the WAN consists of geographically regional communities each of which share a comparatively large overlap in their trees. The GCSN, on the other hand, exhibits subdivisions into groups which are geographically less coherent.

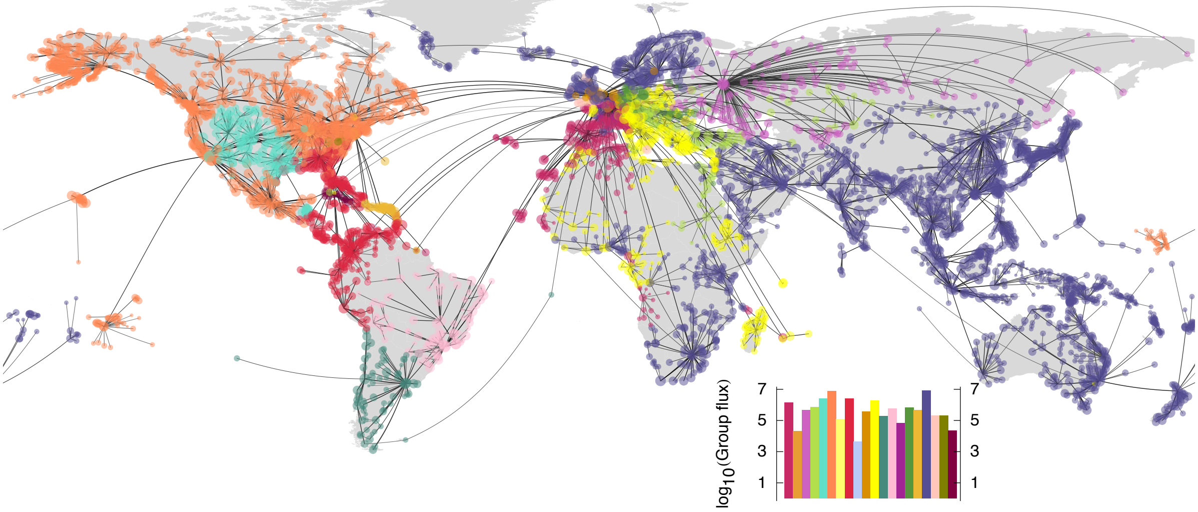

Figure 11: Backbones and effective communities in the worldwide air-transportation network. The map depicts the high-salience connections of the network ($s > \frac{1}{2}$, not necessarily a connected subnetwork) as well as 20 effective communities of nodes based on hierarchical clustering of shortest-path trees using dissimilarity. The bar chart shows the sum of node flux within each group on a logarithmic scale.

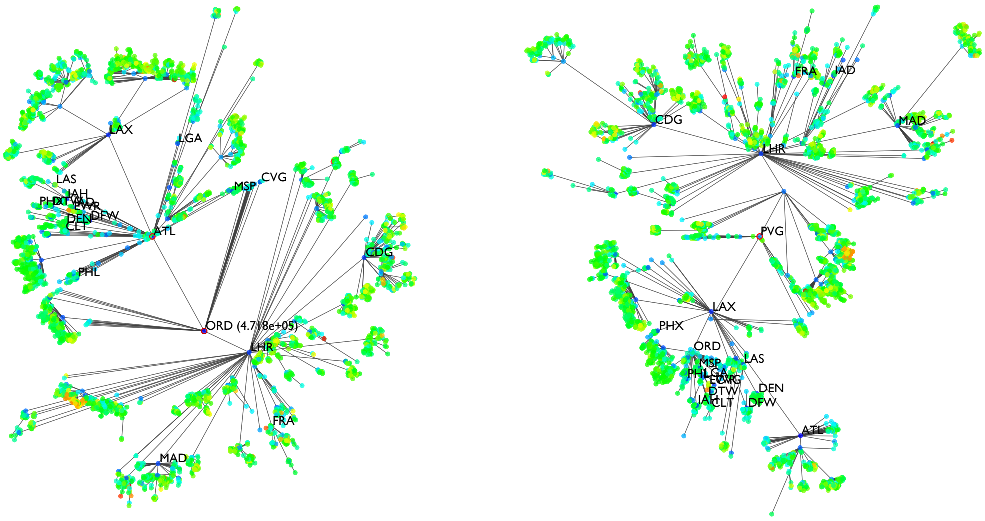

Figure 12: Shortest-path tree dissimilarity in the WAN. The figure shows the same trees as in Fig. 5 (**left:** the airports ATL (Atlanta) and **right:** PVG (Shanghai). The radial distance in these pictures corresponds to the shortest-path-tree dissimilarity (Eq. 17), while node color encodes node capacity (from blue (high capacity) via green and yellow to red).

Bibliography

[1] R Albert, H Jeong, and AL Barabasi. Error and attack tolerance of complex networks. *Nature*, 406(6794):378–382, Jan 2000.

[2] Uri Alon. Network motifs: theory and experimental approaches. *Nature Reviews: Genetics*, 8(6):450–61, June 2007.

[3] AL Barabasi and R Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, Jan 1999.

[4] A Barrat, M Barthelemy, and Alessandro Vespignani. The effects of spatial constraints on the evolution of weighted complex networks. *J Stat Mech-Theory E*, page P05003, Jan 2005.

[5] Stephen P. Borgatti and Martin G. Everett. A Graph-theoretic perspective on centrality. *Social Networks*, 28(4):466–484, October 2006.

[6] D Brockmann, L Hufnagel, and T Geisel. The scaling laws of human travel. *Nature*, 439:462–465, Jan 2006.

[7] D Brockmann and F Theis. Money circulation, trackable items, and the emergence of universal human mobility patterns. *IEEE Pervas Comput*, 7(4):28–35, Jan 2008.

[8] Guido Caldarelli. *Scale-Free Networks: Complex Webs in Nature and Technology*. Oxford University Press, 2007.

[9] Yiping Chen, Gerald Paul, Shlomo Havlin, Fredrik Liljeros, and H. Eugene Stanley. Finding a better immunization strategy. *Phys Rev Lett*, 101(5):058701, Jan 2008.

[10] R Cohen, K Erez, D ben Avraham, and S Havlin. Resilience of the internet to random breakdowns. *Phys Rev Lett*, 85(21):4626–4628, Jan 2000.

[11] R Cohen, S Havlin, and D ben Avraham. Efficient immunization strategies for computer networks and populations. *Phys Rev Lett*, 91(24):247901, Dec 2003.

[12] L Dall'Asta, A Barrat, M Barthelemy, and Alessandro Vespignani. Vulnerability of weighted networks. *J Stat Mech-Theory E*, page P04006, Jan 2006.

[13] Daniel Grady, Christian Thiemann, and Dirk Brockmann. *In preparation*, 2012.

[14] Daniel Grady, Christian Thiemann, and Dirk Brockmann. Robust classification of salient links in complex networks. *Nature Communications (submitted)*, 2012.

[15] Roger Guimerà, S. Mossa, A. Turtschi, and Luís A. Nunes Amaral. The worldwide air transportation network: Anomalous centrality, community structure, and cities' global roles. *PNAS*, 102(22):7794–9, May 2005.

[16] Pablo Kaluza, Andrea Koelzsch, Michael T Gastner, and Bernd Blasius. The complex network of global cargo ship movements. *J R Soc Interface*, 7(48):1093–1103, Jan 2010.

[17] M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45(2):167–256, 2003.

[18] M. E. J. Newman and Michelle Girvan. Finding and evaluating community structure in networks. *Physical Review E*, 69(2):26113, 2004.

[19] Filippo Radicchi, José J. Ramasco, and Santo Fortunato. Information filtering in complex weighted networks. *Physical Review E*, 83(4):1–9, April 2011.

[20] Erzsébet Ravasz and Albert-László Barabási. Hierarchical organization in complex networks. *Physical Review E*, 67(2):1–7, February 2003.

[21] M. Angeles Serrano, Marián Boguñá, and Alessandro Vespignani. Extracting the multiscale backbone of complex weighted networks. *PNAS*, 106(16):6483–8, April 2009.

[22] Christian Thiemann, Fabian Theis, Daniel Grady, Rafael Brune, and Dirk Brockmann. The structure of borders in a small world. *PLoS ONE*, 5(11):e15422, 11 2010.

[23] Christian Thiemann, Fabian Theis, Daniel Grady, Rafael Brune, and Dirk Brockmann. The Structure of Borders in a Small World. *PLoS ONE*, 5(11):e15422, November 2010.

[24] M. Tumminello, T. Aste, T. Di Matteo, and R. N. Mantegna. A tool for filtering information in complex systems. *PNAS*, 102(30):10421–6, July 2005.

[25] Alessandro Vespignani. Predicting the behavior of techno-social systems. *Science*, 325(5939):425–428, Jan 2009.

[26] Huijuan Wang, Javier Martin Hernandez, and Piet Van Mieghem. Betweenness centrality in a weighted network. *Physical Review E*, 77(4):1–10, April 2008.

[27] Olivia Woolley Meza, Christian Thiemann, Daniel Grady, and Dirk Brockmann. *In preparation*, 2012.

[28] Olivia Woolley-Meza, Christian Thiemann, Daniel Grady, Jake Jungbin Lee, Hanno Seebens, Bernd Blasius, and Dirk Brockmann. Complexity in human transportation networks: A comparative analysis of worldwide air transportation and global cargo ship movements. *European Physical Journal B*, 84(4):589–600, 2011.

[29] Zhenhua Wu, Lidia A. Braunstein, Vittoria Colizza, Reuven Cohen, Shlomo Havlin, and H. Eugene Stanley. Optimal paths in complex networks with correlated weights: The worldwide airport network. *Phys. Rev. E*, 74(5):056104, Nov 2006.